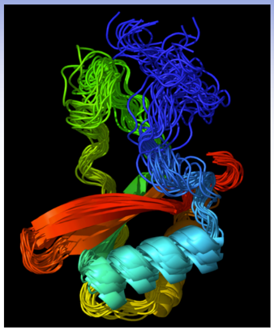
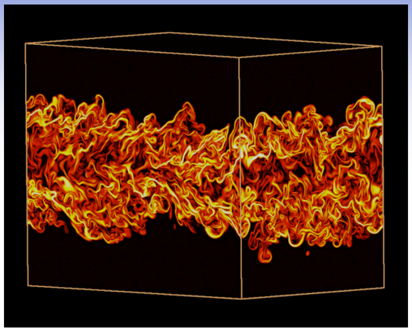
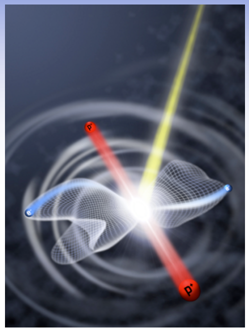
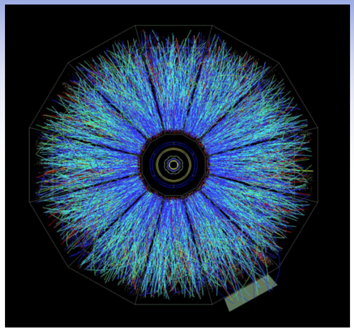
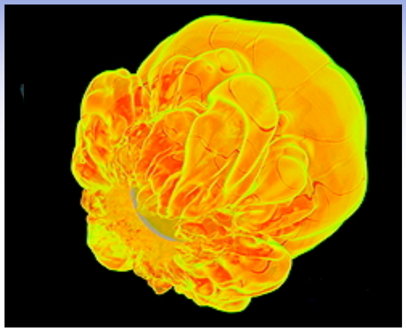






Emerging Application and Algorithm Requirements for Future HPC Systems

July 2008

Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC

BERKELEY LAB

Traditional Sources of Performance Improvement are Flat-Lining

- **New Constraints**
  - 15 years of *exponential* clock rate growth has ended
- **But Moore's Law continues!**
  - How do we use all of those transistors to keep performance increasing at historical rates?
  - Industry Response: #cores per chip doubles every 18 months *instead* of clock frequency!

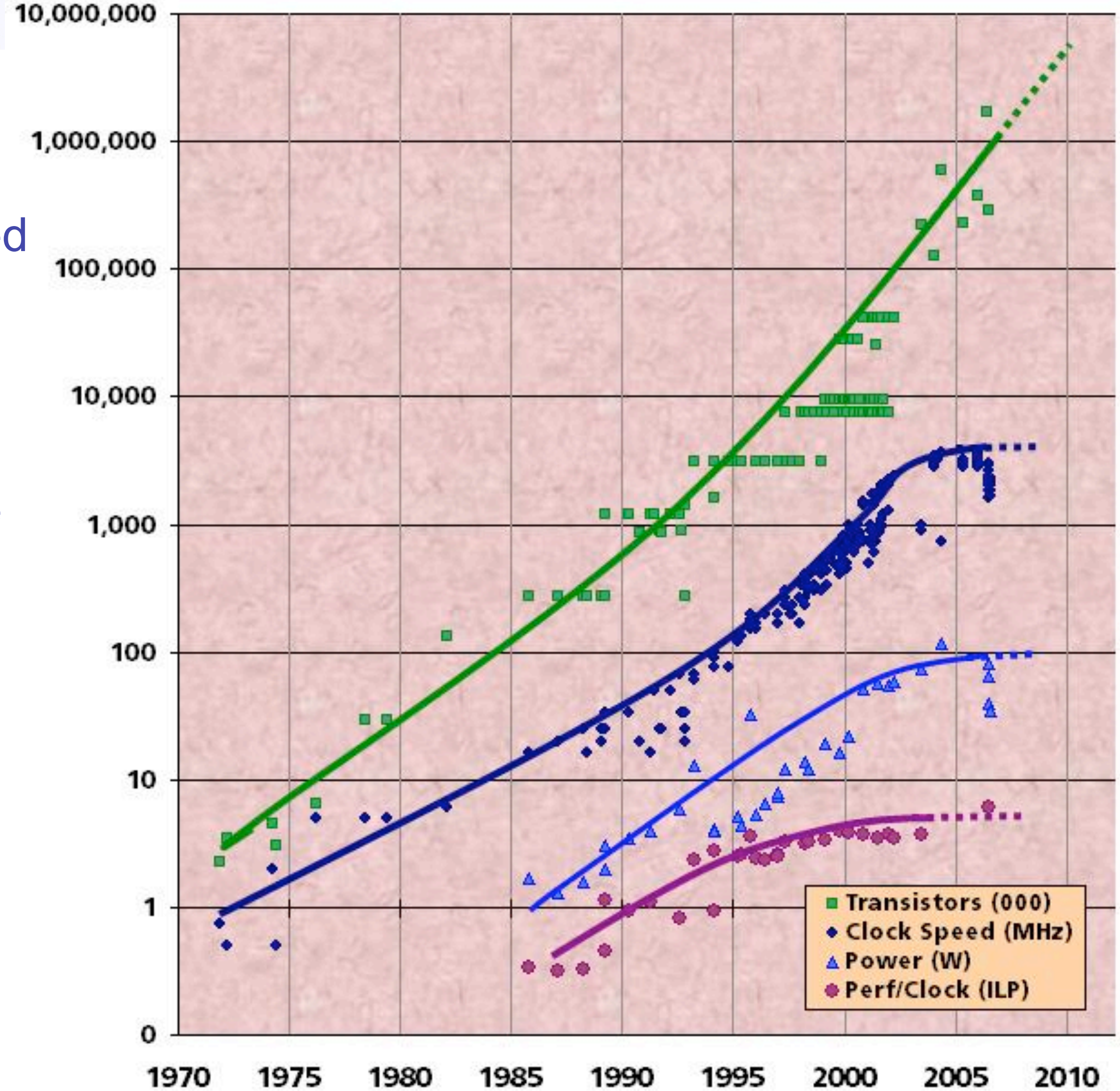
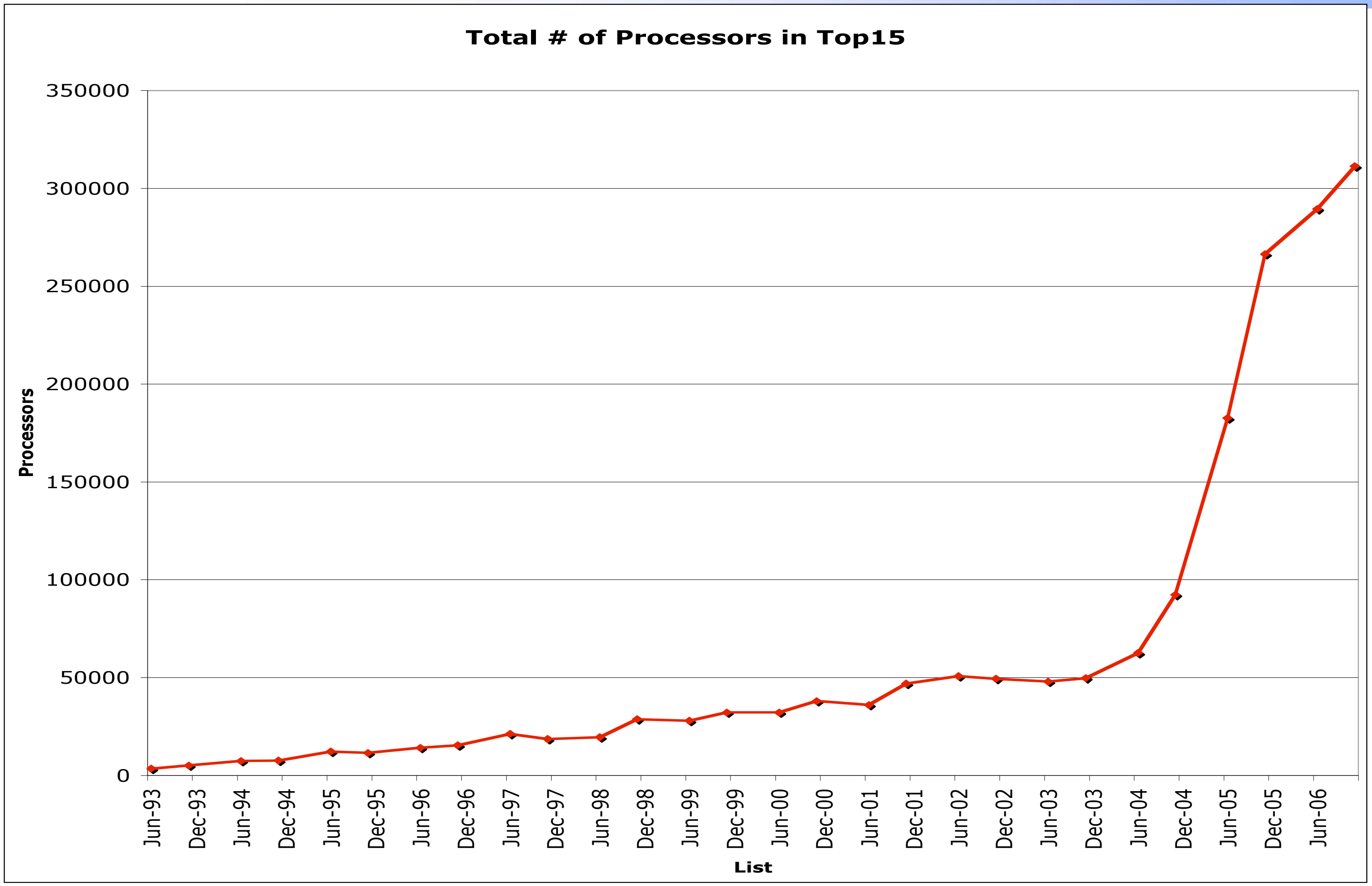


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Growth in HPC System Concurrency



Must ride exponential wave of increasing concurrency for forseeable future!

You will hit 1M cores sooner than you think!

Application Community's Response to Technology Trends

- **Parallel computing has thrived on weak-scaling for past 15 years**
- **Flat CPU performance increases emphasis on strong-scaling**
- **Workload Requirements will change accordingly**
  - Concurency will increase proportional to system scale (3-5x increase over NERSC-5)
  - Timestepping algorithms will be increasingly driven towards implict or semi-implicit stepping schemes
  - Multiphysics/multiscale problems increasingly rely on spatially adaptive approaches such as Berger-Oliger AMR
  - Strong scaling will push applications towards smaller messages sizes – requiring lighter-weight messaging

NERSC Response To Trends

- **Parallel computing has thrived on weak-scaling for past 15 years**
- **Flat CPU performance increases emphasis on strong-scaling**
- **NERSC-6 Benchmarks changed accordingly**
  - Increased concurrency 4x over NERSC-5 benchmarks
  - Input decks emphasize strong-scaled problems
  - Emphasis on implicit methods
  - New AMR benchmark
  - New UPC benchmark

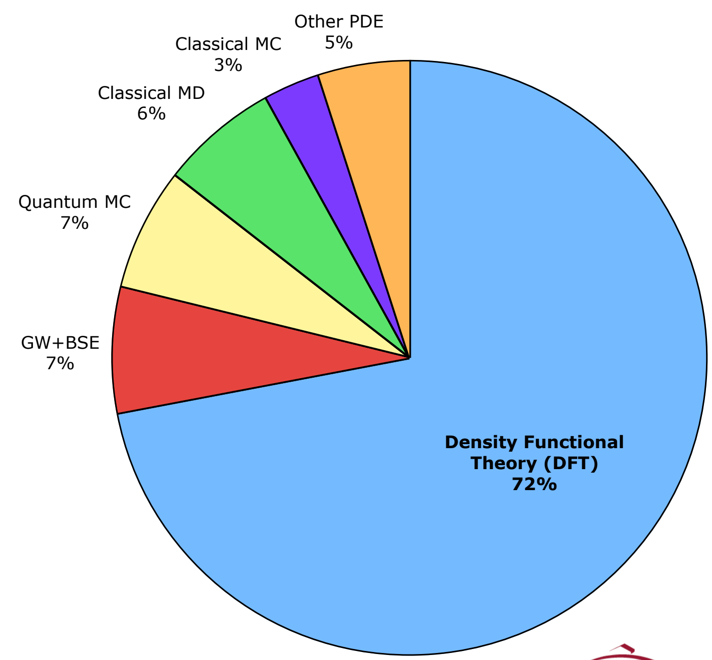
Materials Science

Planewave Density Functional Theory (DFT)

Density Functional Theory (DFT) Algorithm

- **Kohn-Sham formalism for computing electronic structure from first principles (DFT Method)**
  - Most common implementation is based on expanding the quantum wavefunction into plane-wave (fourier) components
  - This is the method employed by VASP, PARATEC, and Qbox

- **Dominant phases of planewave DFT algorithm**
  - **3D FFT**
    - transforming between real space and reciprocal space
    - O(Natoms^2) complexity
  - **Subspace Diagonalization**
    - O(Natoms^3) complexity
  - **Orthogonalization**
    - dominated by BLAS3
    - ~O(Natoms^3) complexity
  - **Compute Non-local pseudopotential**
    - O(Natoms^3) complexity



Andrew Canning

7

Future of Materials Science Codes

- **For smaller atomic systems (~600-1000 atoms)**
  - **BLAS dominates at lower concurrencies**
  - **3D FFT tends to dominate the computation at high concurrency**
    - Due to low computational intensity and small message size (NSF Track-2 bench)
    - Message size can be increased by expending more memory/processor
- **For larger atomic systems (>1k atoms), the $O(N^3)$ complexity of orthogonalization and computing non-local pseudopotential will dominate**
- **For $O(N^3)$ complexity, moving from teraflops to petaflops only gets you from 1k atoms to 4k atoms.**
  - not very impressive given the amount of hardware!
  - Good news is that FLOP rates will be very impressive given increased domination of highly localized BLAS3 operations (eg QBox example)
- ***For this reason, conventional O(N^3) DFT will be increasingly supplanted by O(N) methods for Petaflop scale calculations!***

Anatomy of an O(N) DFT method

(LS3DF as an example)

- **Total energy of a system can be decomposed into two parts**
  - **Quantum mechanical part:**
    - wavefunction kinetic energy and exchange correlation energy
    - Highly localized
    - Computationally expensive part to compute
  - **Classical electrostatic part:**
    - Coulomb energy
    - Involves long-range interactions
    - Solved efficiently using poisson equation even for million atom systems

- **LS3DF exploits localization of quantum mechanical part of calculation**
  - Divide computational domain into discrete tiles and solve quantum mechanical part
  - Solve global electrostatic part (no decomposition)
  - Very little interprocessor communication required! (almost embarrassingly parallel)
  - Result is O(Natoms) complexity algorithm: enables exploration of larger atomic systems as we move to petaflop and beyond.

Lin-Wang Wang

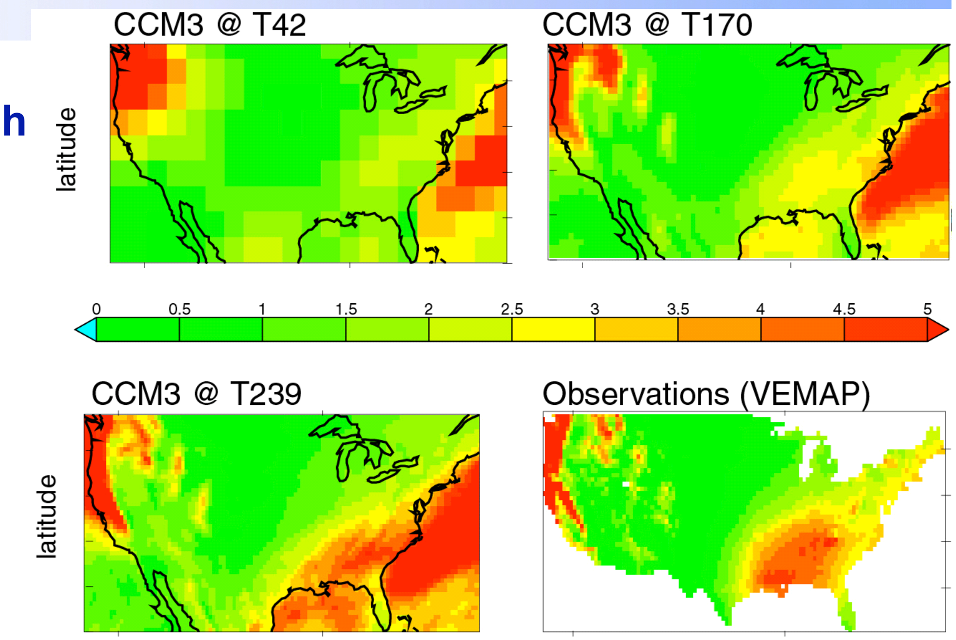
Climate

Cloud System Resolving Climate Simulation

- **Requires *transformational* change in science not feasible using current approach**
  - **The biggest source of climate model errors is poor cloud simulation, *especially tropical convection***
  - At ~1 km horizontal resolution, cloud systems can be resolved

- **DOE Investment in Exascale Computing**
  - Climate change is leading justification for general purpose exascale system
  - Not achievable via extrapolation of current approach
  - **UN WMO Climate Modeling Summit: 1km models are the top priority**

- **Requires substantial code redevelopment to develop cloud-resolving climate model**



Global Cloud System Resolving Climate Modeling

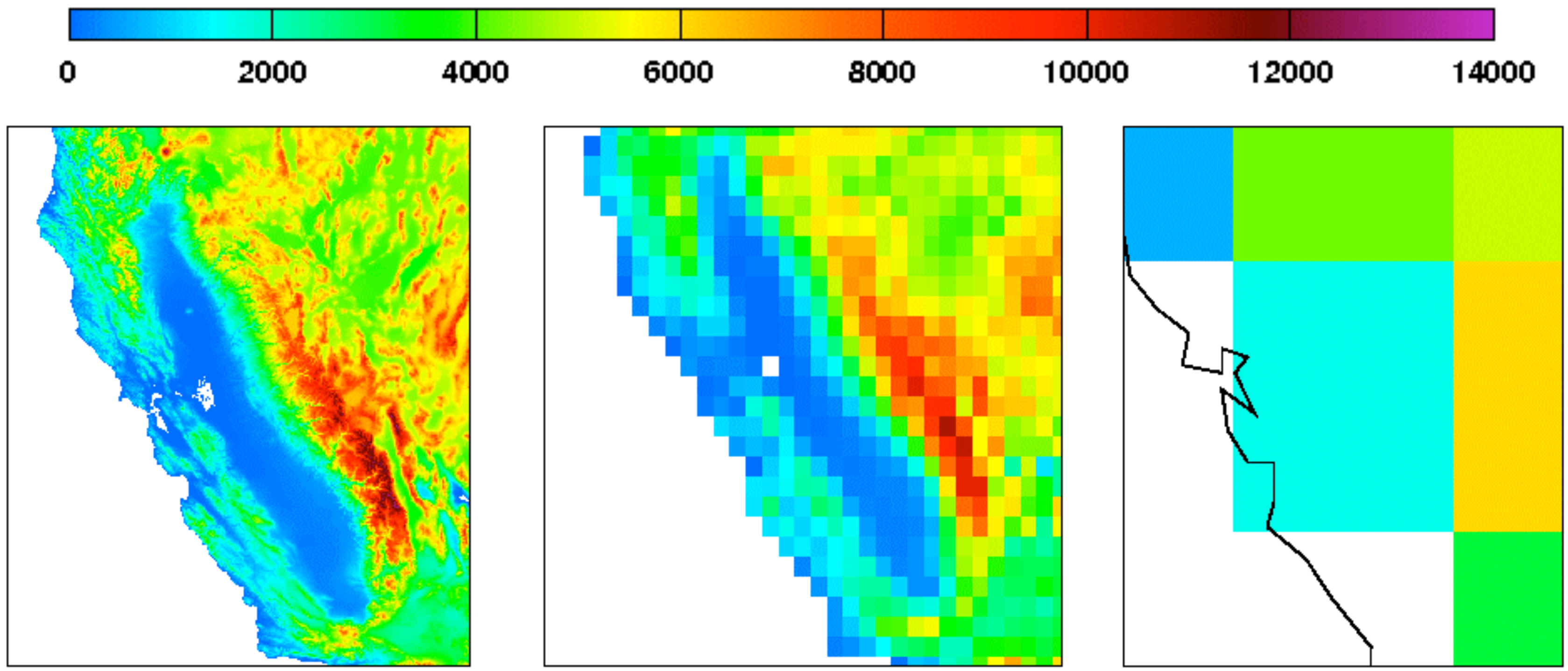
Cloud-scale processses
Well understood → **Meso-scale statistics**
Poorly understood → **Global scale**

This is where parameterization comes in.

Courtesy Prof. David Randall, Colorado State University

The UN WMO cites the need for Cloud Resolving Models as a Top Priority
(cannot be accomplished without 10^7 improvement in computational capability)

Global Cloud System Resolving Models are a Transformational Change



1km
Cloud system resolving models

25km
Upper limit of climate models with cloud parameterizations

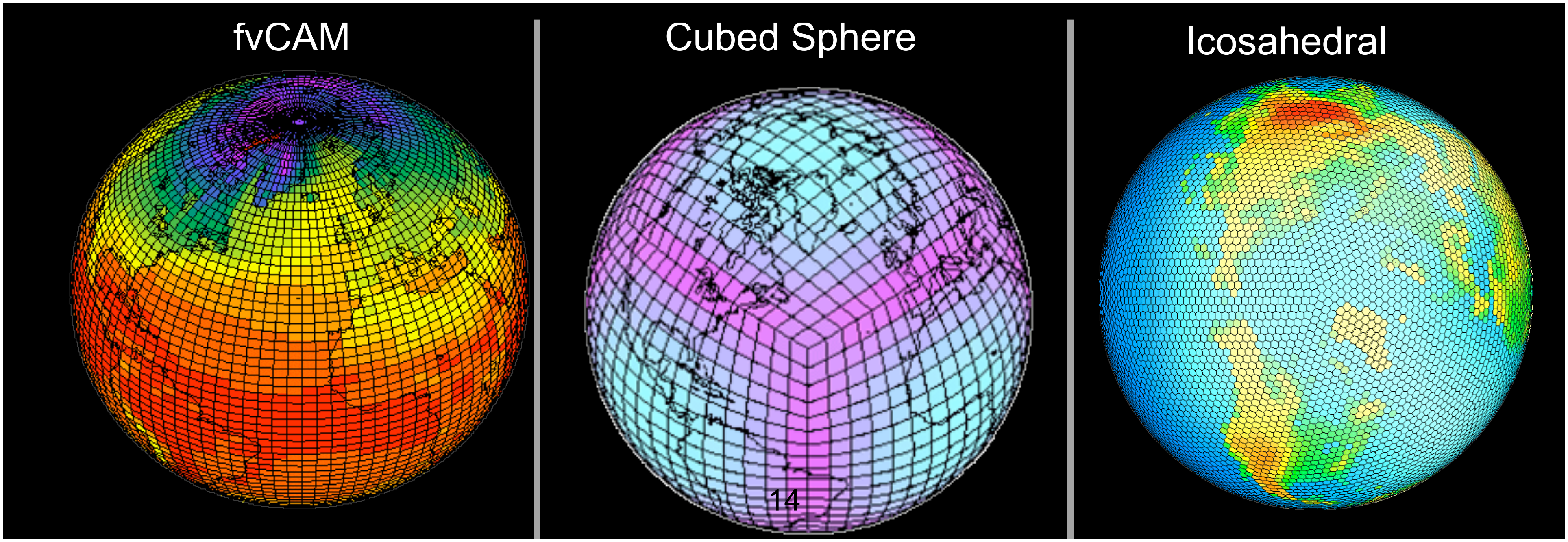
200km
Typical resolution of IPCC AR4 models

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Climate Model

New Approaches for Massive Parallelism

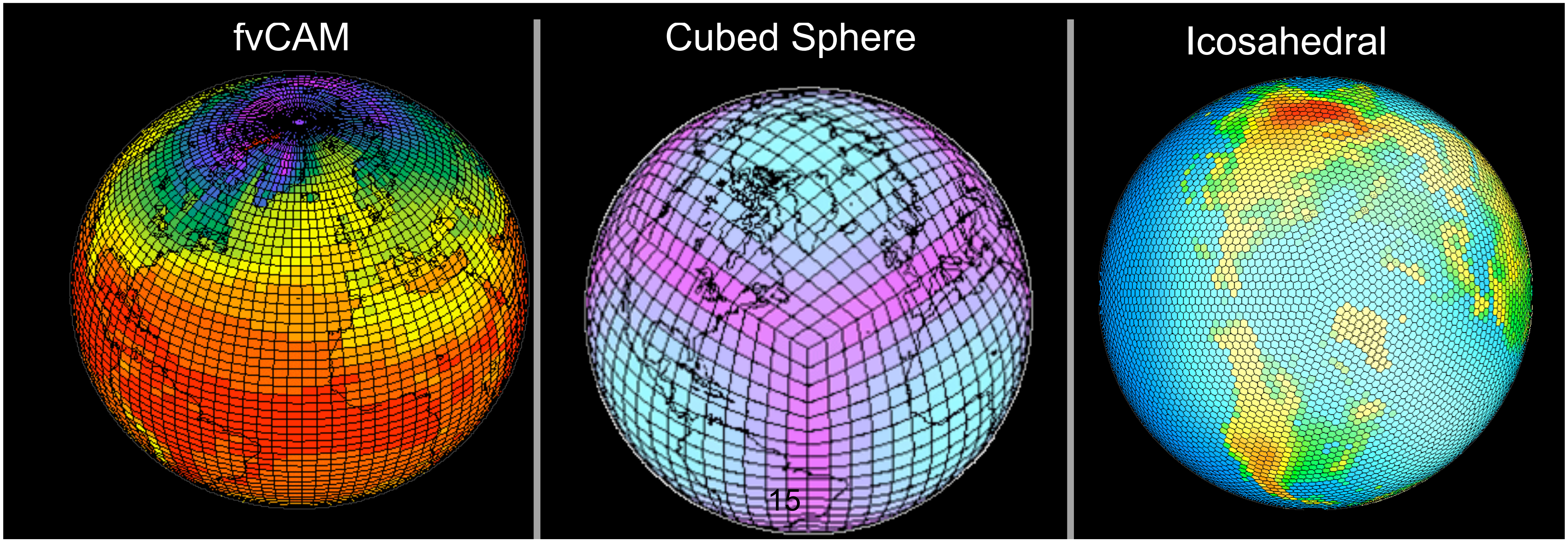
- **Existing Latitude-longitude based algorithm advection algorithm breaks down significantly before 1km scale!**
  - **Grid cell aspect ratio at the pole is 10000!**
  - **Advection time step is problematic at this scale**
- **Ultimately requires new discretization for atmosphere model**
  - **Must expose sufficient parallelism to exploit power-efficient design**
  - **Partner with CSU/Randall Group to use the Icosahedral Code**
  - **Uniform cell aspect ratio across globe**



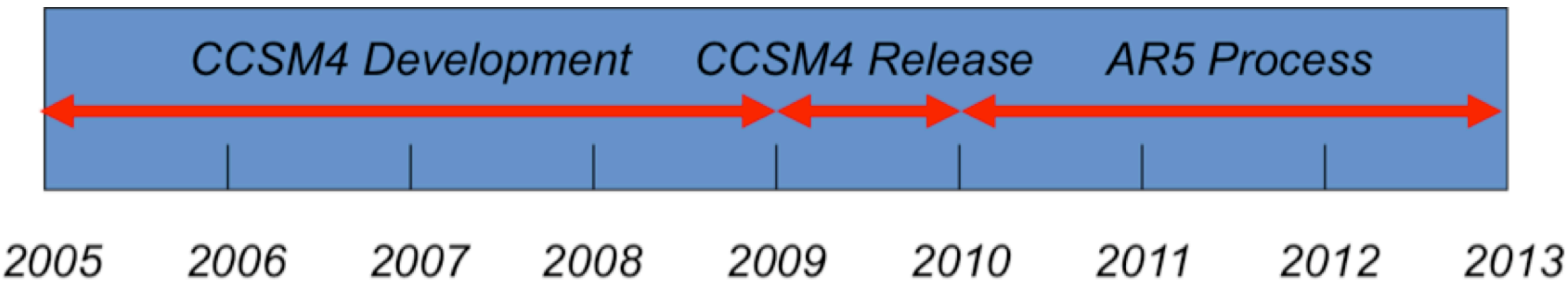
Requirements: 1km Climate Model

Must maintain 1000x faster than real time for practical climate simulation

- **~2 million horizontal subdomains**
- **100 Terabytes of Memory**
  - **5MB memory per subdomain**
- **~20 million total subdomains**
  - **500Mflops sustained per domain**
  - **Nearest-neighbor communication 250GB/s**
- ***NERSC supports projects developing these new discretizations***
  - ***GFDL Cubed Sphere, CSU Icosahedral model***



IPCC AR5 Timeline Coincident with NERSC-6



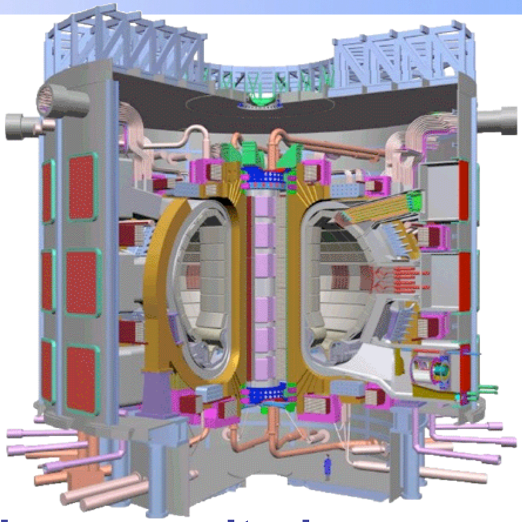
“The carbon cycle version of CCMS4 will include the additional bio-geochemistry, indirect aerosol and land ice components, and the short-term climate simulations will have considerably enhanced atmosphere resolution and, potentially, include the chemistry component. [The] carbon cycle CCSM 4 will be a factor of about five times the CCSM 3 in computing cost. . . . Doing all the proposed IPCC AR 5 runs will stretch the CCSM computing resources to the absolute limit.”

Peter R. Gent:

CCSM4 Implementation Plan

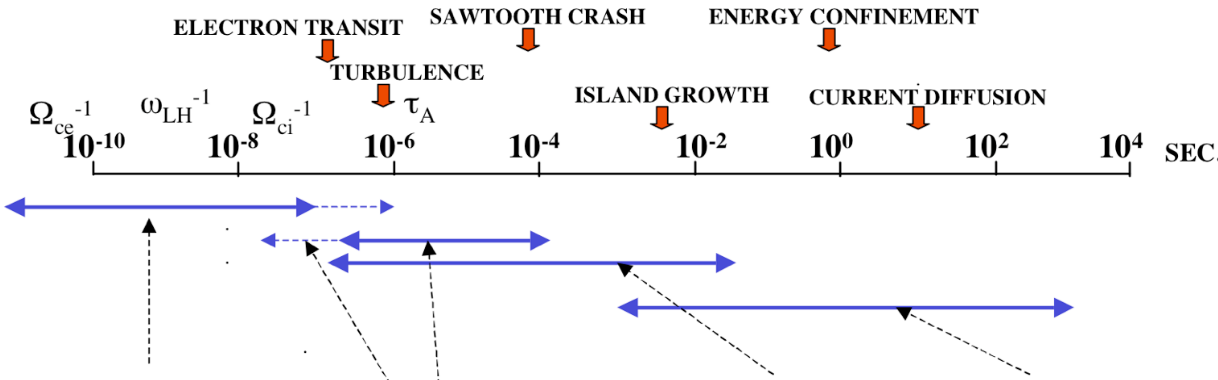
Fusion

Fusion: Impact of ITER



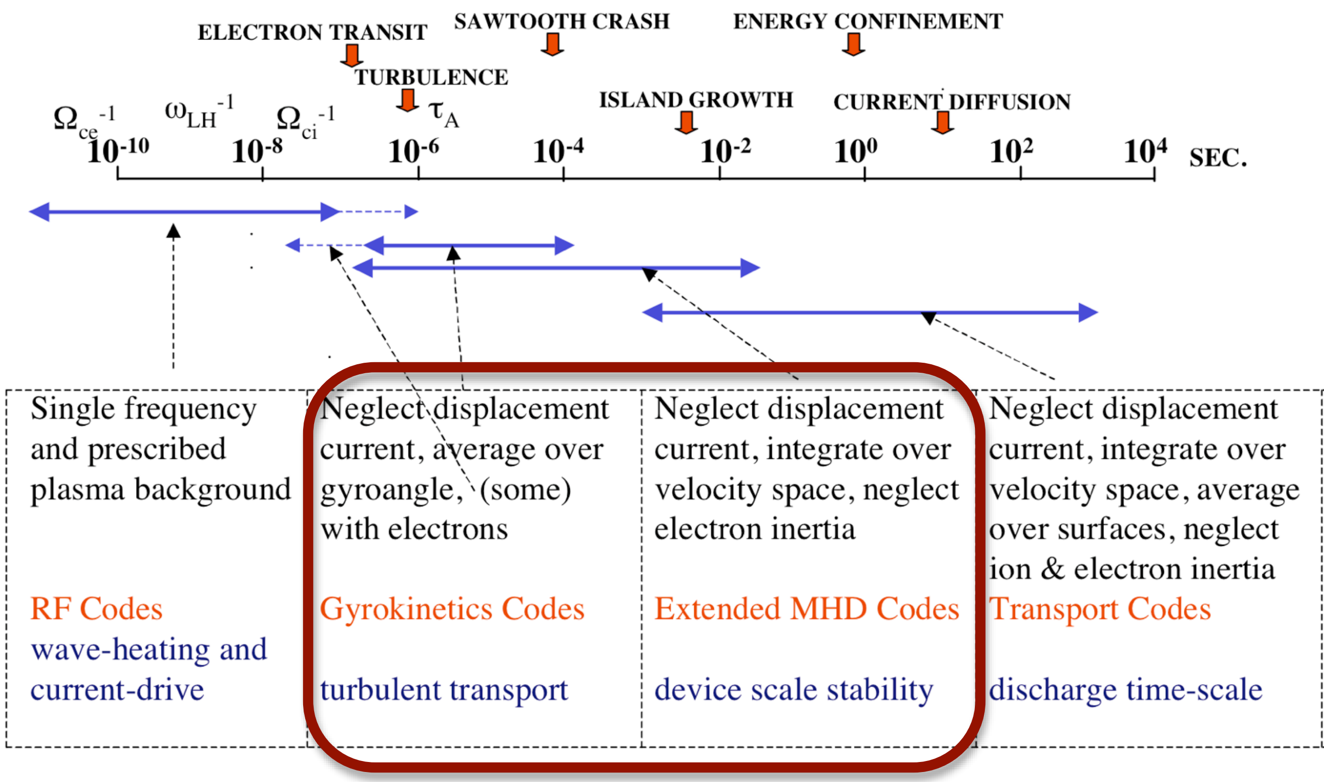
- **Fusion science has been dominated by scaling up first-principles models of specific phenomena**
- **ITER development requires full-device modeling capability by 2012**
  - **For shot planning and device control**
  - **Requires Code-coupling, Multi-scale multiphysics**
  - **Uncontrolled discharge could damage $12B device!**
- **Requires new code and algorithms to span 12 orders magnitude (Keyes/Jardin)**
  - AMR to cover 3 orders of magnitude (time and resolution)
  - Implicit solvers to cover 4 orders magnitude (time)
  - Increased parallel scaling to cover another 3 orders magnitude
  - 2 orders magnitude from higher order elements
- **These codes are still in development (and need a platform to support development)**
  - SciDAC developing pairwise code coupling
  - FSP will focus on broader coupling for full device modeling capability

Fusion Time and Length Scales



Single frequency and prescribed plasma background	Neglect displacement current, average over gyroangle, (some) with electrons	Neglect displacement current, integrate over velocity space, neglect electron inertia	Neglect displacement current, integrate over velocity space, average over surfaces, neglect ion & electron inertia
RF Codes	Gyrokinetics Codes	Extended MHD Codes	Transport Codes
wave-heating and current-drive	turbulent transport	device scale stability	discharge time-scale

Fusion Time and Length Scales



- **Gyrokinetic and MHD codes dominate workload**
  - **GTC (10%) & GEM (11%) PIC codes dominate Gyrokinetic Codes**
  - **M3D (10%) & NIMROD (12%) dominate Extended MHD Codes**

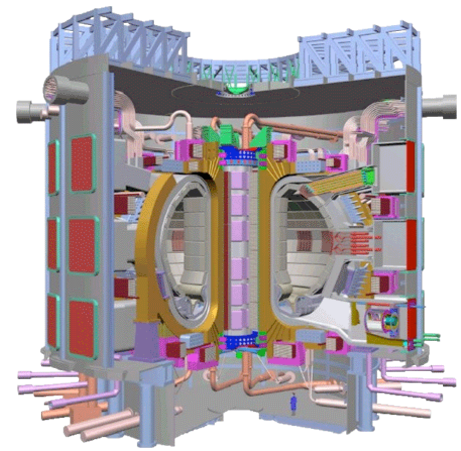
Emerging Workload Requirements

- **Applying computation only where needed**
  - **AMR: multiscale/multiresolution physics**
    - Load balancing issues
    - Locality constraints for prolongation and restriction
    - Many very small (oddly-sized) messages for interconnect
  - **Sparse Matrix: Don't compute on non-zeros**
    - Very small messages sizes and load balance issues
- **Emerging issues with existing applications**
  - **Implicit Methods**
    - Vector inner product required by Krylov subspace algorithms is hampered by latency-bound fast global reductions at massive parallelism
  - **Climate Models**
    - When science that depends on parameter studies and ensemble runs, capacity and capability are intimately linked!
- **I/O Intensive workloads**
  - Growth in experimental and sensor data processing

Scaling Fusion Simulations Up to ITER

name	symbol	units	CDX-U	DIII-D	ITER
Field	B_0	Tesla	0.22	1	5.3
Minor radius	a	meters	.22	.67	2
Temp.	T_e	keV	0.1	2.0	8.
Lundquist no.	S		1×10^4	7×10^6	5×10^8
Mode growth time	$\tau_A S^{1/2}$	s	2×10^{-4}	9×10^{-3}	7×10^{-2}
Layer thickness	$aS^{-1/2}$	m	2×10^{-3}	2×10^{-4}	8×10^{-5}
zones	$N_R\times N_\theta\times N_\phi$		3×10^6	5×10^{10}	3×10^{13}
CFL timestep	$\Delta X/V_A$ (Explicit)	s	2×10^{-9}	8×10^{-11}	7×10^{-12}
Space-time pts			6×10^{12}	1×10^{20}	6×10^{24}

10^{12} needed (explicit uniform baseline)



International **T**hermonuclear **E**xperimental **R**eactor

in Cadaraches, France, operational by 2017

Slides from David Keys -- Altered some for NERSC6

How to Increase Efficiency?

Hardware

- Increased processor speed and efficiency
- Increased concurrency

Software

- Higher-order discretizations
  - Same accuracy can be achieved with many fewer elements
- Flux-surface following gridding
  - Less resolution required along than across field lines
- Adaptive gridding
  - Zones requiring refinement are <1% of ITER volume and resolution requirements away from them are ~10^2 less severe
- Implicit solvers
  - Mode growth time 9 orders longer than Alfven-limited CFL

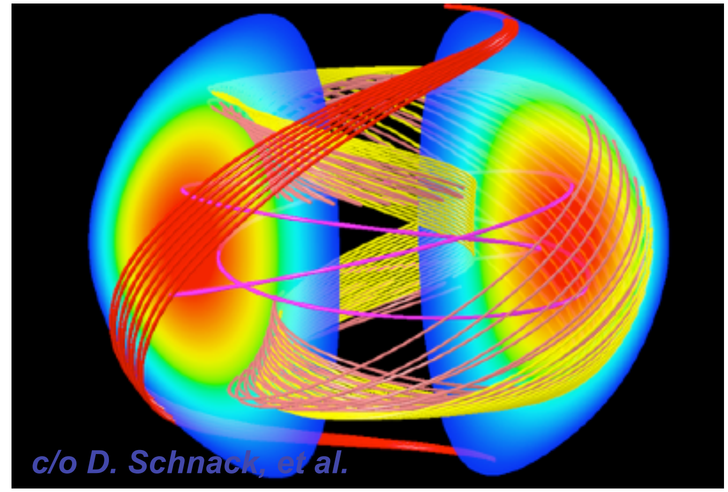
Illustrations from Computational MHD

- **M3D code (Princeton)**
  - **multigrid replaces block Jacobi/ASM preconditioner for optimality**
  - **new algorithm callable across *Ax=b* interface**

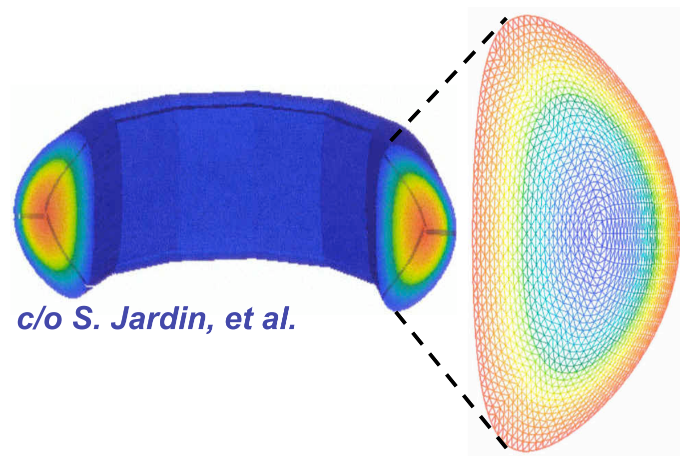
- **NIMROD code (General Atomics)**
  - **direct elimination replaces PCG solver for robustness**
  - **scalable implementation of old algorithm for *Ax=b***

Computational MHD

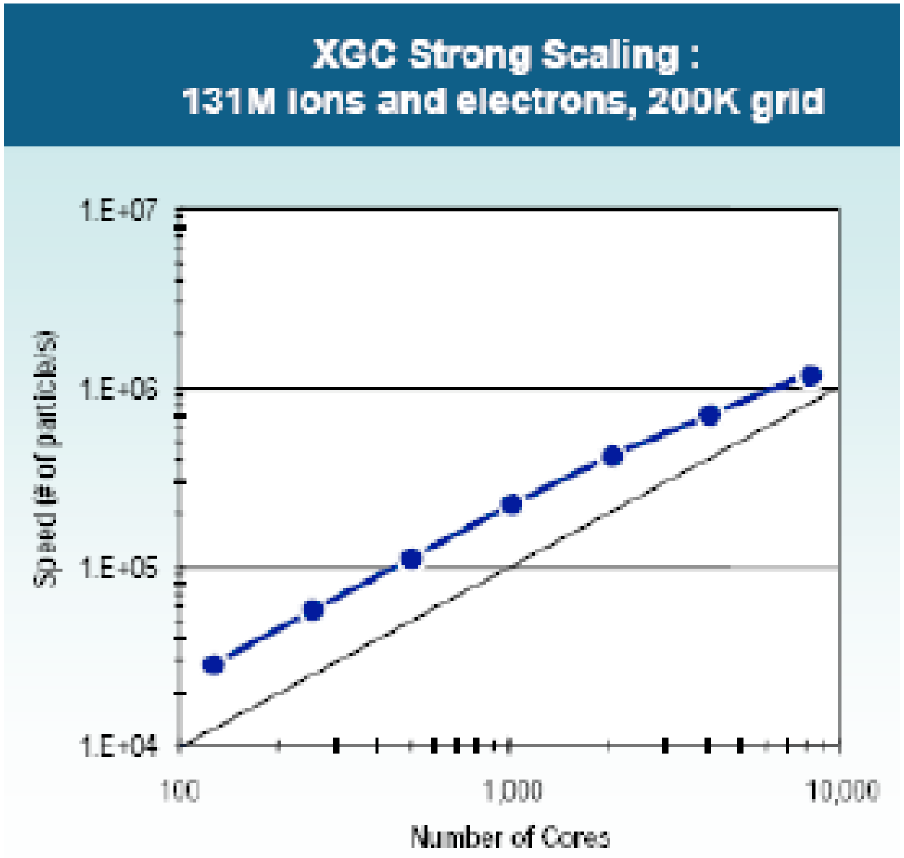
- **NIMROD code: Direct Elim. for robustness**
  - Fourier transforms in toroidal direction
  - High-order finite elements in 2D poloidal crossplanes
  - Sequence of complex, nonsymmetric linear systems with 10K-100K unknowns in 2D (>90% exe. Time)
  - Uses SuperLU (parallel sparse direct solver benefits from efficient support of very small messages sizes)



- **M3D code: multigrid for optimality**
  - Finite differences in toroidal direction
  - Unstructured mesh, hybrid FE/FD discretization with C0 elements in 2D poloidal crossplanes
  - Sequence of real scalar systems (>90% exe. Time)
  - algebraic multigrid (AMG) from Hypre (multigrid benefits from good support of fine-grained messaging)

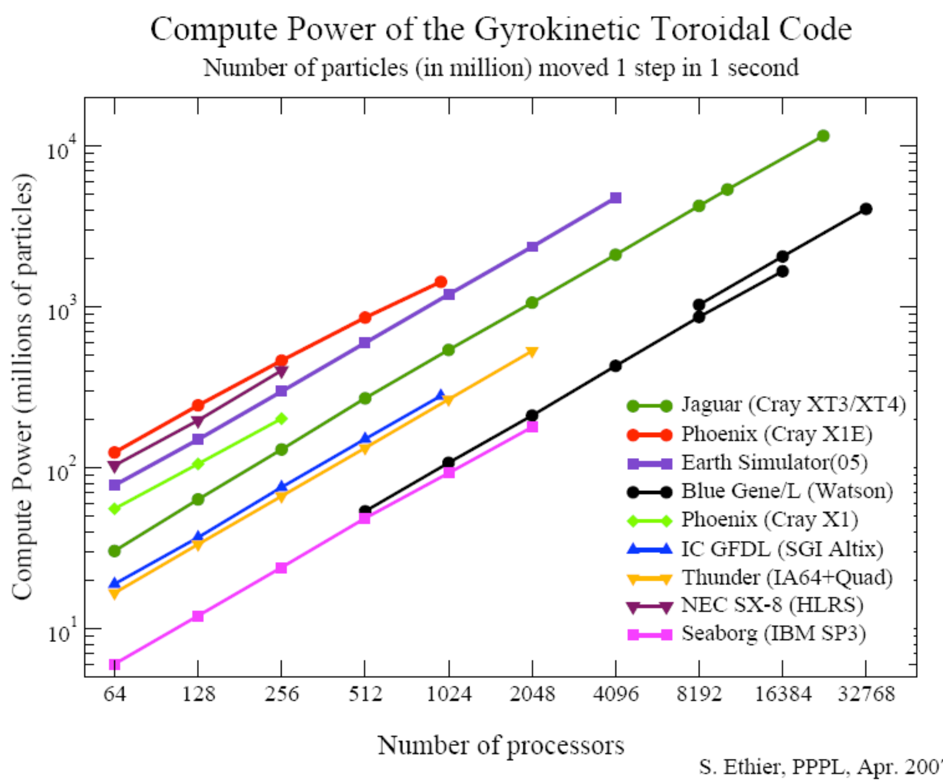


Scaling of PIC Codes



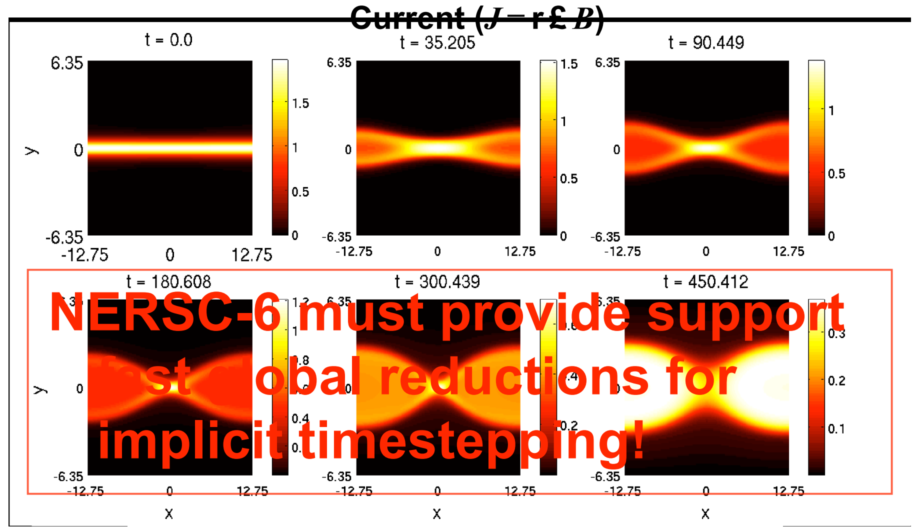
FSP example
(C.S. Chang)

SciDAC example
(S. Ethier)

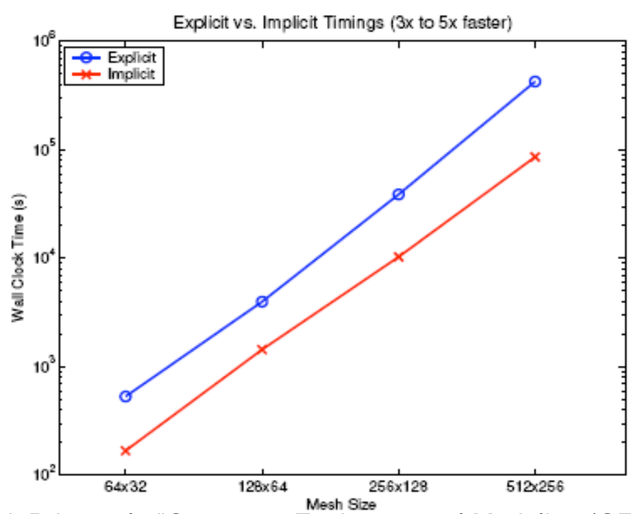


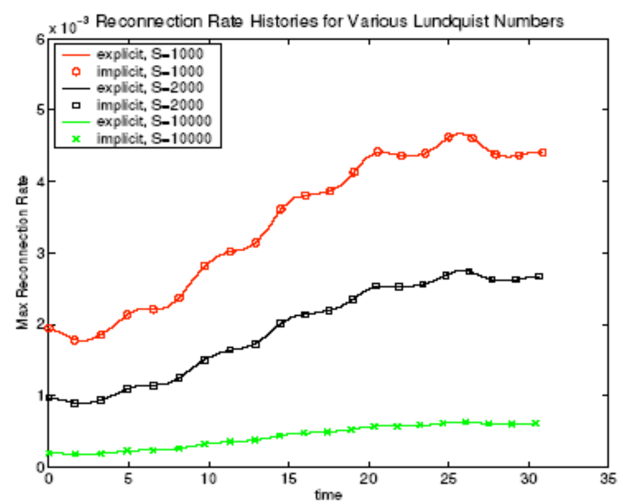
Resistive MHD: Nonlinear Implicit Model

- ***Magnetic reconnection***: the breaking and reconnecting of oppositely directed magnetic field lines in a plasma, replacing hot plasma core with cool plasma, halting the fusion process
- Replace explicit timestepping with implicit Newton-Krylov from SUNDIALS with factor of ~5× in execution time



NERSC-6 must provide support fast global reductions for implicit timestepping!



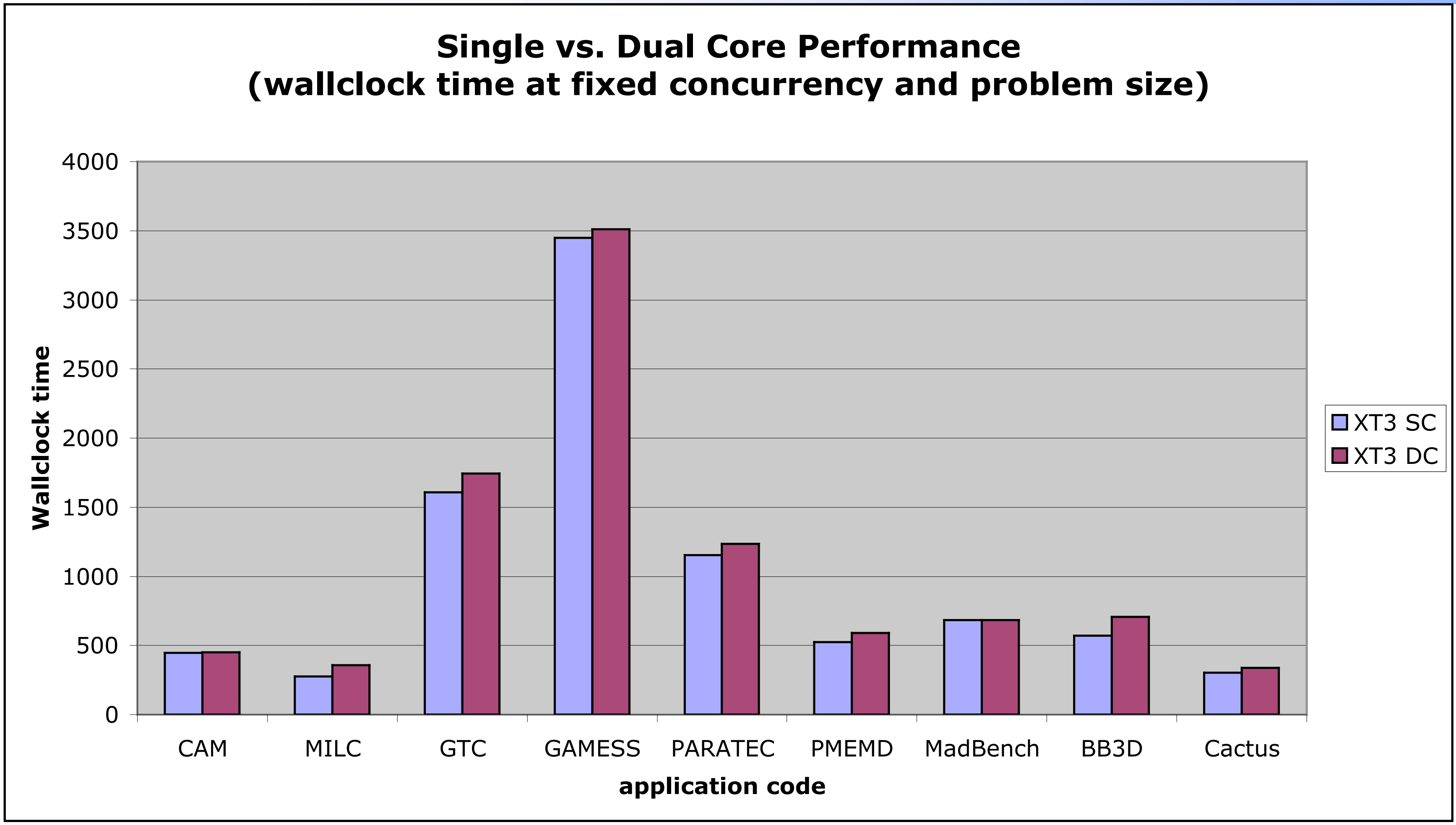


J. Brin et al., “Geospace Environmental Modeling (GEM) magnetic reconnection challenge,” **J. Geophys. Res.** 106 (2001) 3715-3719.

c/o D. Reynolds, et al.

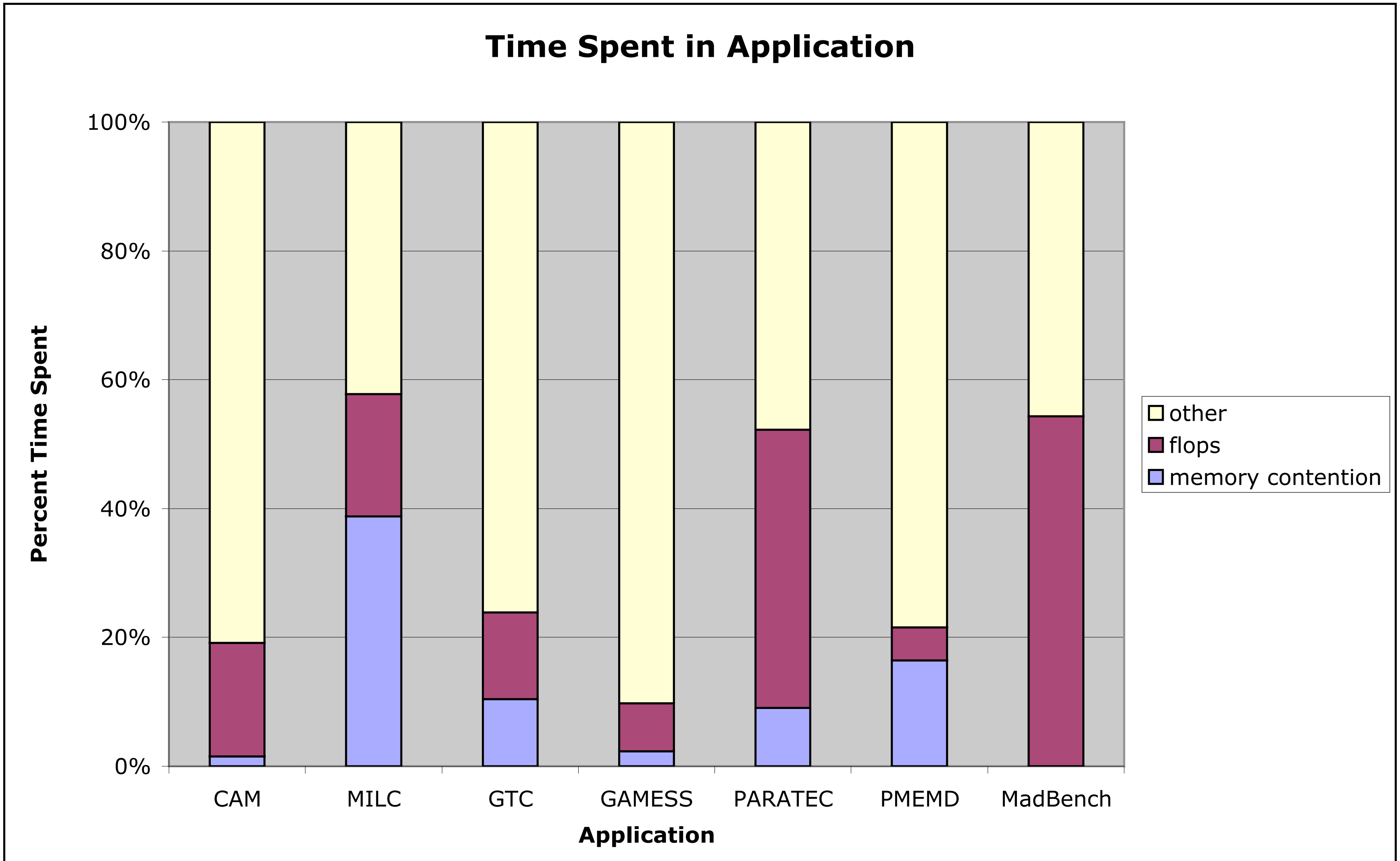
Memory Bandwidth and Interconnect

Sensitivity to Memory Bandwidth

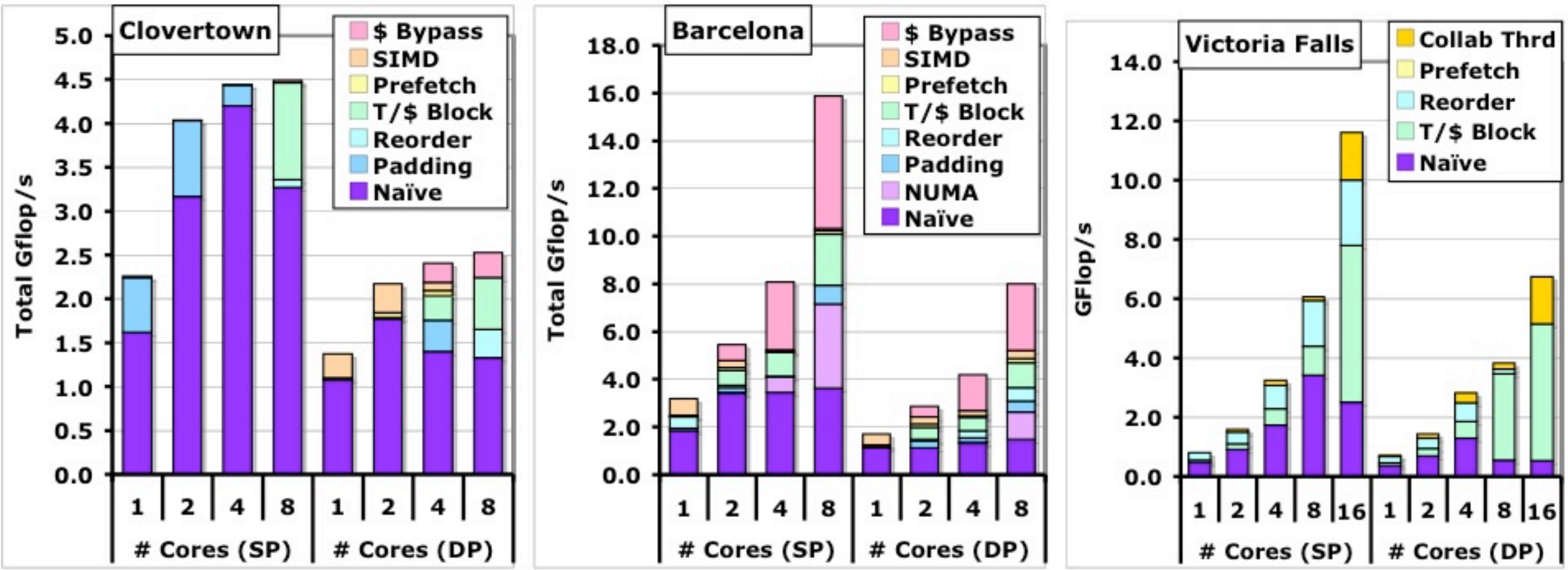


Poor compiler performance makes applications underutilize mem bandwidth
Result: relatively insensitive to halving memory bandwidth

Time spent dominated by "other" (mostly latency stalls)



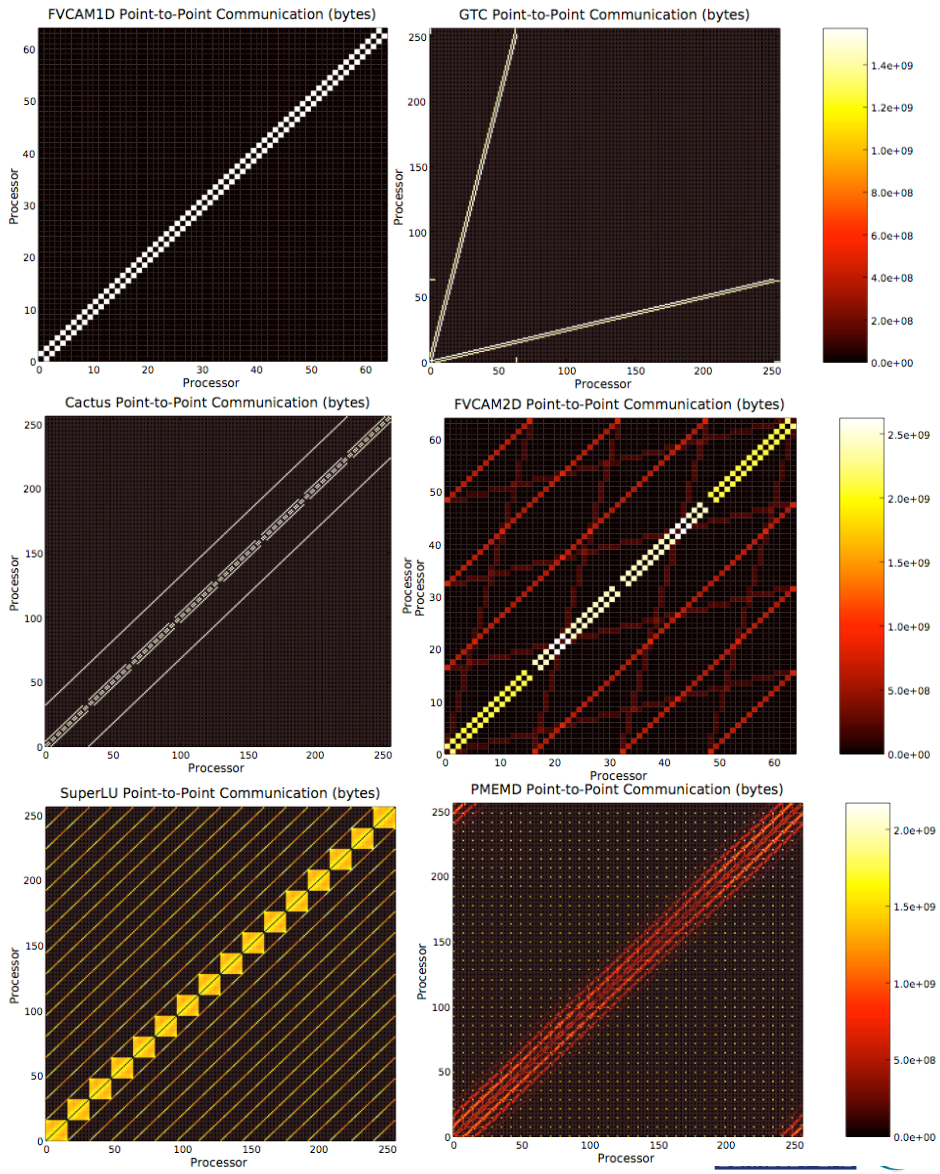
Hand-Tuned Kernels Can Reach Peak (and BW ceiling)



Interconnect Design Considerations for Massive Concurrency

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

- **Application studies provide insight to requirements for Interconnects (both on-chip and off-chip)**
  - On-chip interconnect is 2D planar (crossbar won't scale!)
  - Sparse connectivity for dwarfs; crossbar is overkill
  - No single best topology
- **A Bandwidth-oriented network for data**
  - Most point-to-point message exhibit sparse topology & bandwidth bound
- **Separate Latency-oriented network for collectives**
  - E.g., Thinking Machines CM-5, Cray T3D, IBM BlueGene/L&P
- **Ultimately, need to be aware of the on-chip interconnect topology in addition to the off-chip topology**
  - Adaptive topology interconnects (HFAST)
  - Intelligent task migration?



Office of Science
U.S. DEPARTMENT OF ENERGY

Interconnects
Need For High Bisection Bandwidth

- **3D FFT easy-to-identify as needing high bisection**
  - Each processor must send messages to all PE's! (all-to-all) for 1D decomposition
  - However, most implementations are currently limited by overhead of sending small messages
  - 2D domain decomposition (required for high concurrency) actually requires sqrt(N) communicating partners *(some-to-some)*
- **Same Deal for AMR**
  - AMR communication is sparse, but by no means is it bisection bandwidth limited